School of Computing and Information Systems The University of Melbourne COMP90049

Knowledge Technologies (Semester 1, 2019) Workshop sample solutions: Week 4

Suppose that we have observed the token lended, and we have a dictionary as follows:

addendum
blenders
commodity
deaden
end
leader
leant
lent
lemonade
pleading

- 1. Which, if any, of the above dictionary entries would be returned using a Neighbourhood Search with a neighbourhood of 1? 2? 3?
 - There aren't any items in the dictionary requiring only a single change from lended.
 - With a neighbourhood size of 2, there is a dictionary entry:
 - leader, by Replacing the n with a, and the second d with r
 - Along with the above, the following are also within a neighbourhood of 3:
 - blenders, by Inserting the b, Replacing the second d with r, and Inserting the s
 - deaden (three Replaces)
 - end (three Deletions)
 - lent (one Replace and two Deletions)
- 2. With respect to the input string lended and the dictionary entry deaden, calculate the following:
 - (a) the Global Edit Distance, using the parameter [m, i, d, r] = [+1, -1, -1, -1]

(a)	ε		1		е		n		d		е		d
ε	0	\leftarrow	-1	\leftarrow	-2	\leftarrow	-3	\leftarrow	-4	\leftarrow	-5	\leftarrow	-6
	 	_		_		_		_				_	
d	-1		-1	\leftarrow	-2	\leftarrow	-3		-2	\leftarrow	-3	\leftarrow	-4
	 	_	\uparrow	_						_			
е	-2		-2		0	\leftarrow	-1	\leftarrow	-2		-1	\leftarrow	-2
	 	_	\uparrow		\uparrow	_		_			\uparrow	_	
a	-3		-3		-1		-1	\leftarrow	-2		-2		-2
	1	_	\uparrow		\uparrow	_	\uparrow	_				_	
d	-4		-4		-2		-2		0	\leftarrow	-1		-1
	 	_	\uparrow	_	\uparrow	_	\uparrow		\uparrow	_			
е	-5		-5		-3		-3		-1		1	\leftarrow	0
	 	_	\uparrow		\uparrow	_			\uparrow		\uparrow	_	
n	-6		-6		-4		-2		-2		0		0

• From the table above, we can observe that the Global Edit Distance is 0, corresponding to the following sequence of operations: Replace, Match, Replace, Match, Match, Replace, which I will abbreviate as rmrmmr. (You can follow along with the highlighted backpointers.)

(b)	ε	1	е	n	d	е	d
ε	0	0	0	0	0	0	0
				7		_	
d	0	0	0	0	1 .	\leftarrow 0	1
					† *		
е	0	0	1 ←	- 0	0	$2 \leftarrow$	1
			↑	\		↑	
a	0	0	0	0	0	1	1
				7		↑ \	
d	0	0	0	0	1 +	\leftarrow 0	2
			_		↑ •		\uparrow
е	0	0	1 ←	- 0	0	$2 \leftarrow$	1
			↑ ×	_		↑	
n	0	0	0	2 *	← 1	1	1

- (b) the Local Edit Distance, using the parameter [m, i, d, r] = [+1, -1, -1, -1]
 - From the table above, we can observe that the Local Edit Distance is 2 (highlighted); there are five equivalent-scoring substring matches that it corresponds to:
 - Align -de- in lended with the first de- in deaden: mm
 - Align -ded with dead-: mmim
 - Align -de- in lended with the second -de- in deaden: mm
 - Align -ende- with -eade-: mrmm
 - Align -en- with -en: mm
- (c) the N-Gram Distance, using n=2
 - We begin by generating the 2-grams of the two strings; I will opt not to use the terminal marker (#) here:
 - lended: le, en, nd, de, ed
 - deaden: de, ea, ad, de, en
 - Recall that the N-Gram Distance is defined as follows:

$$D(s,t) = |G_n(s)| + |G_n(t)| -2 \times |G_n(s) \cap G_n(t)|$$

- Here we have 5 2-grams in lended, as well as 5 in deaden. Also, the two sets share 2 2-grams: de and en. (Note that we don't double-count the des in deaden, because there is only a single de in lended)
- Consequently, the 2-gram Distance is $5 + 5 2 \times 2 = 6$
- 3. Find the best approximate match (or matches, if there are ties) in the dictionary for the string lended, based on the following methods; consider different parameters where necessary:
 - (a) the Global Edit Distance
 - Using the above scoring parameter, the most similar dictionary entries are blenders (+2) and leader (+2)
 - You might like to try some other parameter setting(s), to see if they give different results.
 - (b) the Local Edit Distance
 - Using the above scoring parameter, the best dictionary entry is blenders (+5)
 - In this case, changing the parameter is unlikely to result in a different answer. (Why?)
 - (c) the N-Gram Distance
 - If we are using n is 2 and not padding with #, the best dictionary entry is end, with a 2-Gram Distance of 3.
 - You might find that adding the padding characters or changing n will give different results.
 - (d) Soundex

- The Soundex code of lended is 1533.
- None of the dictionary entries have this exact code; however, if we permit one mismatch in the Soundex code (as in Neighbourhood Search with a neighbourhood of 1), then the best matches are commodity (c533), leant (153), lent (153), and lemonade (1553)